In *Working Notes of the Workshop on 'Bayesian and Causal Networks: From Inference to Data Mining,' 12th European Conference on Artificial Intelligence (ECAI-2000)*, Berlin, Germany, 22 August 2000

Learning Bayesian Network Parameters from Small Data Sets: Application of Noisy-OR Gates

Agnieszka Oniśko [1] and **Marek J. Druzdzel** [2] and **Hanna Wasyluk** [3]

Abstract. Existing data sets of cases can significantly reduce the knowledge engineering effort required to parameterize Bayesian networks. Unfortunately, when a data set is small, many conditioning cases are represented by too few or no data records and they do not offer sufficient basis for learning conditional probability distributions. We propose a method that uses Noisy-OR gates to reduce the data requirements in learning conditional probabilities. We test our method on HEPAR II, a model for diagnosis of liver disorders, whose parameters are extracted from a real, small set of patient records. Diagnostic accuracy of the multiple-disorder model enhanced with the Noisy-OR parameters was around 6% better than the accuracy of the plain multiple-disorder model and 10% better than the single-disorder diagnosis model.

1 INTRODUCTION

Bayesian networks [13] (also called belief networks) are acyclic directed graphs modeling probabilistic dependencies among variables. The graphical part of a Bayesian network reflects the structure of a problem, while local interactions among neighboring variables are quantified by conditional probability distributions. Bayesian networks have earned the reputation of being powerful tools for modeling complex problems involving uncertain knowledge. They have been employed in practice in a variety of fields, including engineering, science, and medicine with some models reaching the size of hundreds of variables.

A major difficulty in applying Bayesian network models to practical problems is the effort that goes in model building, i.e., obtaining the model structure and the numerical parameters that are needed to fully quantify it. The complete conditional probability distribution table (CPT) for a binary variable with n binary predecessors in a Bayesian network requires specification of 2^n independent parameters. For a sufficiently large n, eliciting 2^n numbers from a domain expert may be prohibitively cumbersome. One of the main advantages of Bayesian networks over other schemes for reasoning under uncertainty is that they readily combine existing frequency data with expert judgment within their probabilistic framework. When sufficient amount of data is available, they can be used to learn both the structure and the parameters of a Bayesian network model [2, 14, 16]. The existing learning methods are theoretically sound and are guaranteed to produce very good results given sufficiently large data sets. However, in case of small data sets, quite typical in practice, learned models can be of lesser quality.

The focus of this paper is learning CPTs in Bayesian network models from small data sets given an existing network structure. The problem encountered in small data sets is that many conditioning cases are represented by too few or no data records and they do not offer sufficient basis for learning conditional probability distributions. Learning CPTs amounts essentially to counting data records for different conditions encoded in the network. Roughly speaking, prior probability distributions are simply obtained from relative counts of various outcomes for each of the nodes without predecessors. Conditional probability distributions are obtained from relative counts of various outcomes in those data records that fulfill the conditions described by a given combination of the outcomes of the predecessors. While prior probabilities can be learned reasonably accurately from a database consisting of a few hundred records, learning CPTs is hampered by the exponential growth of their size. In cases where there are several variables directly preceding a variable in question, individual combinations of their values may be very unlikely to the point of being absent from the data file.

A CPT offers a complete specification of a probabilistic interaction that is powerful in the sense of its ability to model any kind of interaction among a node Y and its parents $X_1, \ldots, X_n$. However, the limited size of the data set makes this precision illusory, as many of the CPT entries will be based on few records, undermining the very purpose of a full specification. We propose enhancing the process of learning the CPTs from data by combining the data with structural and numerical information obtained from an expert. Given expert's indication that an interaction in the model can be approximated by a Noisy-OR gate [7, 13], we first estimate the Noisy-OR parameters for this gate. Subsequently, in all cases of a small number of records for any given combination of parents of a node, we generate the probabilities for that case as if the interaction was a Noisy-OR gate. At the same time, the learned distribution is smoothed out by the fact that in all those places where no data is available to learn it, it is reasonably approximated by a Noisy-OR gate. Noisy-OR distributions approximate CPTs using fewer parameters and learning distributions with fewer parameters is in general more reliable [6]. Some applications of the Noisy-OR gates in medical Bayesian models have already been recorded in the past [8, 15].

We test our approach on HEPAR II, a Bayesian network model for diagnosis of liver disorders consisting of 73 nodes. The parameters of HEPAR II are learned from a data set of 505 patient cases. We show that the proposed method leads to an improvement in the quality of the model as measured by its diagnostic performance.

The remainder of this paper is structured as follows. Section 2 introduces the Noisy-OR gate. Section 3 describes our data set and

[1] Białystok University of Technology, Institute of Computer Science, Białystok, 15–351, Poland, aonisko@ii.pb.bialystok.pl

[2] Decision Systems Laboratory, School of Information Sciences, University of Pittsburgh, Pittsburgh, PA 15260, U.S.A., marek@sis.pitt.edu

[3] The Medical Center of Postgraduate Education, and Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Marymoncka 99, Poland, hwasyluk@cmkp.edu.pl

our model. Section 4 illustrates the structural modifications that we performed on the model in order to apply our method. Section 5 explains the details related to obtaining the Noisy-OR parameters. Finally, Section 6 compares diagnostic accuracy of a model learned using the direct CPT method to models whose parameters are learned using our method.

2 THE NOISY-OR GATE

Some types of conditional probability distributions can be approximated by canonical interaction models that require fewer parameters. Very often such canonical interactions approximate the true distribution sufficiently well and can reduce the model building effort significantly.

One type of canonical interaction, widely used in Bayesian networks is known as Noisy-OR gate [7, 13]. Noisy-OR gates are usually used to describe the interaction between n causes $X_1, X_2, \ldots, X_n$ and their common effect Y.[4] The causes X_i are each assumed to be sufficient to cause Y in absence of other causes and their ability to cause Y is assumed independent of the presence of other causes.

The simplest and most intuitive canonical model is a binary Noisy-OR gate [13], which applies when there are several possible causes $X_1, X_2, \ldots, X_n$ of an effect variable Y, where (1) each of the causes X_i has a probability p_i of being sufficient to produce the effect in the abs ence of all other causes, and (2) the ability of each cause being sufficient is independent of the presence of other causes. The above two assumptions allow us to specify the entire conditional probability distribution with only n parameters $p_1, p_2, \ldots, p_n$. p_i represents the probability that the effect Y will be true if the cause X_i is present and all other causes X_j, $j \neq i$, are absent. In other words,

$$p_i = \Pr(y|\overline{x}_1, \overline{x}_2, \ldots, x_i, \ldots, \overline{x}_{n-1}, \overline{x}_n) \tag{1}$$

It is easy to verify that the probability of y given a subset $\mathbf{X}_p$ of the X_is which are present is given by the following formula:

$$\Pr(y|\mathbf{X}_p) = 1 - \prod_{i:X_i \in \mathbf{X}_p} (1 - p_i) \tag{2}$$

This formula is sufficient to derive the complete CPT of Y conditional on its predecessors $X_1, X_2, \ldots, X_n$.

Henrion [7] proposed an extension of the binary Noisy-OR gate for situations where the effect variable can be true even if all its causes are false and called this extended model a *leaky Noisy-OR* gate. Leaky Noisy-OR is applicable to the situations where a model does not capture all possible causes of Y. Arguably, almost all situations encountered in practice belong to this class. This can be modeled by introducing an additional parameter p_0, called the *leak probability*, the combined effect of all unmodeled causes of Y.

$$p_0 = \Pr(y|\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n) \tag{3}$$

p_0 represents the probability that the effect Y will occur spontaneously, i.e., in absence of any of the causes that are modeled explicitly.

Figure 1 shows an example of a Noisy-OR gate, around a liver disease *Steatosis*.

[4] Throughout this paper, upper case letters (e.g., Y) and indexed upper-case letters (e.g., X_i) will stand for random variables. Lower case letters will denote their outcomes (e.g., x_1 is an outcome of a variable X_1). In case of binary random variables, the two outcomes will be denoted by lower case and negated lower case (e.g., the two outcomes of a variable X will be denoted by x and $\overline{x}$). Bold upper case letters (e.g., $\mathbf{X}$) will denote sets of variables.

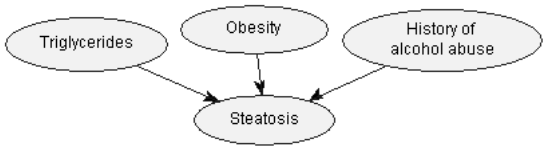

Figure 1. An example of the noisy-Or gate.

Each of *Triglycerides*, *Obesity*, and *History of alcohol abuse* can cause *Steatosis* by itself, although their influence is probabilistic. *Steatosis* can be also caused by some unmodeled factors, which are captured by a leak probability.

In the leaky Noisy-OR gate p_i ($i \neq 0$) no longer represents the probability that X_i causes Y given that all the other causes are absent, but rather the probability that Y is present when X_i is present and all other explicit causes (all the X_j's such that $j \neq i$) are absent.

Let p_i' be the probability that Y will be true if X_i is present and every other cause of Y including unmodeled causes are absent. The more references can be found in [4].

$$1 - p_i' = \frac{1 - p_i}{1 - p_0} \tag{4}$$

From here, we have

$$p_i = p_i' + (1 - p_i')p_0 \tag{5}$$

It follows that the probability of Y given a subset $\mathbf{X}_p$ of the x_i which are present is given in the leaky Noisy-OR gate by the following formula:

$$\Pr(Y|\mathbf{X}_p) \quad = \quad 1 - (1 - p_0) \prod_{i:x_i \in \mathbf{X}_p} \frac{1 - p_i}{1 - p_0}$$

Díez [3] proposed an alternative way of eliciting the parameters of a leaky Noisy-OR gate, which amounts essentially to asking the expert for the parameters p_i' as defined by Eq.(4). The difference between the two proposals has to do with the leak variable. While Henrion's parameters p_i assume that the expert's answer includes a combined influence of the cause in question and the leak, Díez's parameters p_i' explicitly refer to the mechanism between the cause in question and the effect with the leak absent. Conversion between the two parameters is straightforward using equation Eq.(5). If the Noisy-OR parameters are learned from data, Henrion's definition is more convenient, as the observed frequencies include the leak, which is always present by definition.

Two extensions of the binary Noisy-OR gate to nodes including multiple outcomes have been proposed [3, 17] of which we followed the former in our work. Due to lack of space, we refer the reader to the original articles for the details of these extensions.

3 THE HEPAR II MODEL AND DATA

The HEPAR II project [10, 11] aims at applying decision-theoretic techniques to diagnosis of liver disorders. It is a successor of the HEPAR project [1, 18], conducted at the Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences in collaboration with physicians at the Medical Center of Postgraduate Education in Warsaw. The HEPAR system was designed for gathering and processing clinical data of patients with liver disorders and, through its diagnostic capabilities, reducing the need for hepatic biopsy. An integral part of the HEPAR system is its database,

created in 1990 and thoroughly maintained since then at the Gastroentorogical Clinic of the Institute of Food and Feeding in Warsaw. The current database contains over 800 patient records and its size is still growing. Each hepatological case is described by over 200 different medical findings, including patient self-reported data, results of physical examination, laboratory tests, and, finally, a histopathologically verified diagnosis.

One of the assumptions made in the database that was available to us is that every patient case is ultimately diagnosed with only one liver disorder. In other words, the data set assumed that all disorders were mutually exclusive. This assumption led us to the development of a single-disorder diagnosis model. We elicited the structure of the model (i.e., we selected variables from the data set and established dependencies among them) based on medical literature and conversations with our domain expert, a hepatologist Dr. Hanna Wasyluk (third author) and two American experts, a pathologist, Dr. Daniel Schwartz, and a specialist in infectious diseases, Dr. John N. Dowling, from the University of Pittsburgh. We estimate that elicitation of the structure took approximately 40 hours with the experts, of which roughly 30 hours were spent with Dr. Wasyluk and roughly 10 hours spent with Drs. Schwartz and Dowling. This includes model refinement sessions, where previously elicited structure was reevaluated in a group setting. The model is essentially a causal Bayesian network involving a subset of variables included in the HEPAR database. The most recent single-disorder diagnosis version of the model [12], consists of 66 feature nodes and one disorder node covering, in addition to the hepatologically healthy state, 9 mutually exclusive liver disorders: toxic hepatitis, reactive hepatitis, hyperbilirubinemia, chronic hepatitis (active and persistent), steatosis, primary biliary cirrhosis (PBC), and cirrhosis (compensate and decompensate).

The numerical parameters of the model, i.e., the prior and conditional probability distributions, were extracted from the HEPAR database. The data used to extract the numerical parameters contained 505 patient records. All continuous variables were discretized by our expert. One of the assumptions that we used in learning the model parameters was that missing values for discrete finding variables corresponded to state *absent* (e.g., a missing value for *Jaundice* was interpreted as *absent*). In case of continuous variables, a missing value corresponded to a normal value, elicited from the expert (e.g., a missing value for *Bilirubin* was interpreted as being in the range of 0–5), which included the typical value for a healthy patient.

Given a patient's case, i.e., values of some of the modeled variables, such as symptoms or test results, the model derives the posterior probability distribution over the possible liver disorders. This probability distribution can be directly used in diagnostic decision support.

4 STRUCTURAL CHANGES TO THE HEPAR II MODEL

In order to be able to apply parametric probability distributions, such as Noisy-OR gates, in learning the network parameters, we had to restructure the network in such a way that various nodes express either binary propositions or various grades of intensity of some quantity. The disorder node in the single-disorder diagnosis version of the HEPAR II model is a categorical variable with 10 outcomes that is not suitable for a parametric probability distribution. One way of preparing the structure for these distributions is by breaking the disorder node into separate nodes for each of the disorders. This modification addresses two problems: it relaxes the assumption of mutual exclusivity of disorders and makes the nodes more amenable to parametric quantification.

We have concentrated the structural changes on the disorders. We split the disorder node with its 9 mutually exclusive disorders into five binary nodes (*Toxic hepatitis*, *Reactive hepatitis*, *Steatosis*, *Hyperbilirubinemia* and *PBC*) and two nodes with three outcomes each (*Chronic hepatits* and *Cirrhosis*). The feature nodes that we originally modeled as causes/effects of the single *Liver Disorder* variable were broken down into several groups, specific for each of the 9 disorders. As far as the data used in learning the parameters are concerned, we worked with 66 features and 505 records in the database. The resulting model consisted of 73 nodes (66 feature nodes and 7 disorder nodes).

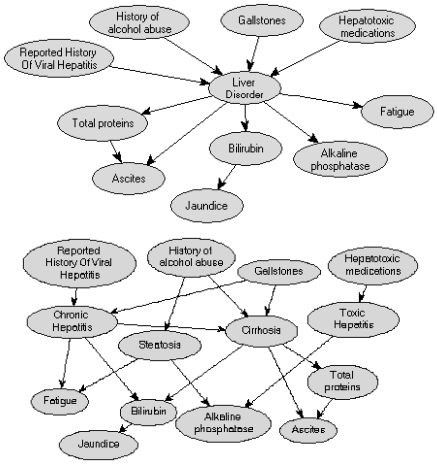

Figure 2. A simplified fragment of the HEPAR II network: single-disorder diagnosis (top) and multiple-disorder diagnosis (bottom) version.

Figure 2 shows a simplified fragment of both models and gives an idea of the structural changes performed in the transition from the single-disorder to the multiple-disorder versions of the model. In particular, the models share each of the four risk factors (*Reported history of viral hepatitis*, *History of alcohol abuse*, *Gallstones*, and *Hepatotoxic medications*) and six symptoms and test results (*Fatigue*, *Jaundice*, *Bilirubin*, *Alkaline phosphatase*, *Ascites*, and *Total proteins*). The single *Liver disorder* node is replaced by four disorder nodes (*Chronic hepatitis*, *Steatosis*, *Cirrhosis*, and *Toxic hepatitis*).

A positive side-effect of our structural changes is that they have decreased the number of numerical parameters required to quantify the model. The main difference between the models is that some of the four new disorder nodes are not connected with some of the risk factors and symptoms. While adding a node increases the number of parameters, it is compensated by removing an outcome of a variable and removing some arcs. The latter especially leads to a logarithmic decrease in the size of a CPT. Our transformation has resulted in a significant reduction of the number of numerical parameters necessary to quantify the network. This, in turn, increased the average number of records for each combination of parents in a CPT. Indeed, the multiple-disorder version of the model required only 1,488 parameters (we counted $\mu = 87.8$ data records per parent combination) compared to the 3,714 parameters ($\mu = 16.8$ data records per parent combination) needed for the single-disorder version of the model. Figure 3 shows the distribution over the number of data records per parent combination for the single-disorder and the multiple-disorder models. We can see that over 50% of the conditional probability distributions in the single-disorder model contained zero records. In the

multiple-disorder model this number is dramatically smaller — only 0.5% of all cases involved zero records and there is quite a high proportion of conditional probability distributions for which tens of records were available. With an increase in the average number of records per parent combination, we can expect the quality of the model parameters to improve.

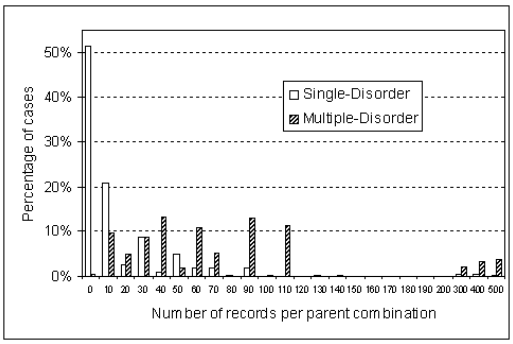

Figure 3. Distribution over the number of data records per parent combination for the single-disorder and the multiple-disorder models.

Unfortunately, structural changes also introduced certain problems. The fact that we used a data set in which each patient record had a single-disorder diagnosis placed us before a difficulty in assessing CPTs of nodes that had several disorder nodes as parents — there were no records in the database for conditions involving combinations of various disorders. We applied a simple solution, in which we included in the calculation all records that described the disorders present in the condition. For example (see Figure 2), when computing the conditional probability distribution of the node *Fatigue* given presence of both *Chronic hepatitis* and *Steatosis*, we used both: records that were diagnosed as *Chronic hepatitis* and records that were diagnosed as *Steatosis*. This amounted to averaging the effect of various disorders. We also tried taking the maximum effect of all disorders present in the condition, with a very modest improvement in performance. Another limitation of the HEPAR data that has a serious implication on our work is that mutual exclusivity of disorders did not allow us to extract dependencies among disorders. Hepatology often deals with disorders that are consequences of the previous disorders, e.g., a chronic liver disorder implies hepatic fibrosis which can further cause cirrhosis. In the future we plan to model and quantify these dependencies by combining data with expert judgment.

5 OBTAINING PARAMETERS FOR NOISY-OR GATES

For each combination of a node and its parents in the multiple-disorder version of the HEPAR II model we verified with the expert whether the interaction could be approximately modeled as an Noisy-OR gate. The expert has identified a total of 27 nodes (from among the total of 62 nodes with parents) that can be reasonably approximated by Noisy-OR gates. Testing the Noisy-OR assumption for each of the gates with the expert was quite straightforward once the expert has understood the concept of independence of causal interaction. When deciding whether an interaction can be approximated by a Noisy-OR gate we followed criteria proposed by Díez [5].

Each of the such identified Noisy-OR gates was subject to the following learning enhancement. Whenever there were sufficiently many records for a given condition, we used these records to learn a corresponding element of the CPT. We have also obtained the Noisy-OR parameters of the gate and then used these parameters to generate conditional probability distributions for those cases that lacked records. The remaining conditional probability distributions for this gate were learned from the data records. Effectively, the complete CPT, once learned, was a general CPT with only some elements generated using the Noisy-OR assumption. The assumption that we made was that a general conditional probability table will fit the actual distribution better than a Noisy-OR distribution. Noisy-OR will fit better than a uniform distribution in those cases when there was not enough data to learn a distribution. In the following two sections we describe two methods of obtaining the Noisy-OR parameters of the gates in question.

5.1 Obtaining Noisy-OR Parameters from Data

We learned the Noisy-OR parameters from data for each of the 27 Noisy-OR gates using equation Eq.(1). We learned the leak parameter using equation Eq.(3). Obtaining the parameters from records that contain a combination of values of parent outcomes would be less reliable, as there would be certainly fewer such records (a conjunction of two events is at most as likely as each of these events in separation). It might be theoretically possible to obtain better estimates of the Noisy-OR parameters by trying to fit these to a larger fragment of a CPT. However, in our initial approach, we used the above simple method.

5.2 Expert Assessments of Noisy-OR Parameters

For each of the 27 Noisy-OR gates identified by the expert, we have also obtained all numerical parameters using direct expert elicitation. There was a total of 153 parameters and the assessment took about 4 hours of the expert time.

Initially, we have posed the expert two types of questions, corresponding to the two theoretical formalizations of the Noisy-OR gate proposed in the literature. The first type of questions focused on the parameters p_i Eq.(1) and was based on Henrion's [7] definition and for the example network fragment in Figure 1 it amounted to:

> What is the probability that obesity results in steatosis when neither triglycerides nor history of alcohol abuse are present?

The second type of questions focused on parameters p_i' Eq.(4) and was based on Díez's [3] definition and it amounted to:

> What is the probability that obesity results in steatosis when no other cause of steatosis is present?

We stumbled across two interesting empirical questions: (1) which of the two definitions is more intuitive for a human expert, and (2) which leads to better quality assessments. While we have not tested (2), in the course of elicitation our expert clearly developed preference for Díez's definition. Using Eq.(5), we subsequently converted the parameters elicited from the expert in Díez's format into Henrion's format, which is the current native format of our software, GeNIe and SMILE.

While we have no objective basis for comparing the quality of expert assessment to the numbers obtained from the data (please note that we evaluated our model using the data, so the best we can say is whether the expert's judgments matched the data or not), we observed a systematic difference between the two: our expert provided usually higher estimates than those learned from the data. An extreme example is shown in Figure 4. Here the causal strengths elicited from the expert are much higher than those learned from the data ($\mu = 0.37$ vs. $\mu = 0.19$ the expert and data respectively).

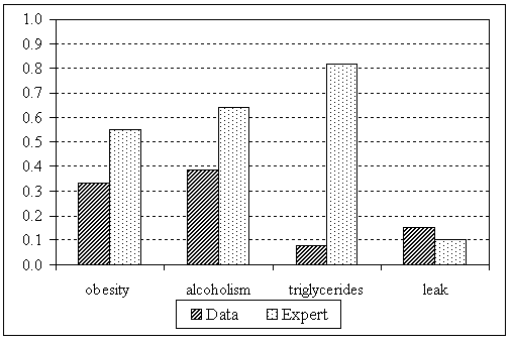

Figure 4. An extreme example of discrepancy between Noisy-OR parameters obtained from the expert and estimated from the data (node *Steatosis*, parents *Triglycerides*, *Alcoholism* and *Obesity*, see Figure 2.)

6 COMPARISON OF DIAGNOSTIC ACCURACY OF THE MODELS

We performed a series of empirical tests of diagnostic accuracy of various versions of the model. In order to make the comparison fair, we used the same data set for learning the parameters of each of the models. Our data set contained 505 patient records classified in 9 different disorder classes. In each case we used the same measure of accuracy: diagnostic performance using the leave-one-out method [9]. Essentially, given n=505 data records, we used $n-1$ of them for learning model parameters and the remaining one record to test the model. This procedure was repeated n times, each time with a different data record. In our tests, we used as observations only those findings that have actually been reported in the data (i.e., we did not use the values that were missing, even though we used their assumed values in learning).

By accuracy we mean the proportion of records that were classified correctly. Whenever we report accuracy within a class, we report the fraction of records within that class that were classified correctly.

6.1 Single- vs. Multiple-disorder Diagnosis Model

Our first empirical test focused on a comparison of the diagnostic performance for the single-disorder and the multiple-disorder models. We were interested in overall performance of the models in terms of classification accuracy (each of the disorders was viewed as a separate class that the program predicted based on the values of all the other variables). This test is very conservative towards the multiple-disorder model, as this is the task for which the single-disorder version of the model was designed. We were interested in both (1) whether the most probable diagnosis indicated by the model is indeed the correct diagnosis, and (2) whether the set of k most probable diagnoses contains the correct diagnosis for small values of k (we chose a "window" of k=1, 2, 3, and 4). Results were for the multiple-disorder version of the model approximately 44% (compared to 42% for the single-disorder version), 59% (57%), 68% (68%), and 77% (78%) for k=1, 2, 3, and 4 respectively. In other words, the most likely diagnosis indicated by the model was the correct diagnosis in 44% of the cases. The correct diagnosis was among the four most probable diagnoses as indicated by the model in 77% of the cases. The performance of both versions of the model was similar, with the multiple-disorder version being slightly more accurate.

In order to gain some insight into when multiple-disorder version of the model is better, we looked at the relationship between the number of records in the database for each class and the diagnostic accuracy within that class.

Figure 5 shows this relationship for the window of size 1 (i.e., the most likely disorder). It is clear that accuracy of both models increases significantly with the number of data records. Another interesting trend is that the multiple-disorder model performed often better than the single-disorder model for those disorders that had many records. This promises a higher diagnostic value of our approach when the available data set is sufficiently large, i.e., when the quality of parameters is high.

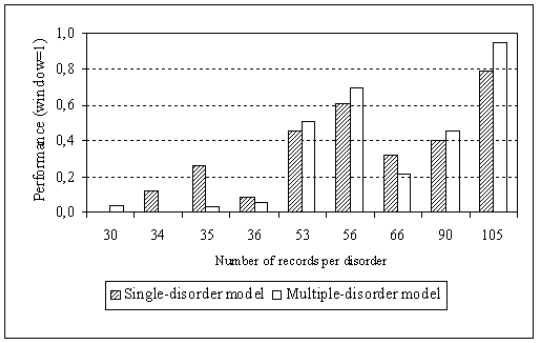

Figure 5. Diagnostic accuracy as a function of the number of disorder cases in the database (class size) for the single and multiple-disorder diagnosis models, window=1.

6.2 Plain CPT Model vs. CPT Smoothed by Noisy-OR Parameters

Our second test aimed at comparing the diagnostic accuracy of the plain multiple-disorder model to the models whose probabilities were smoothed out using the Noisy-OR parameters. Here, we focused on three models: (1) the plain multiple-disorder model (i.e., general CPT) and two models enhanced with: (2) Noisy-OR parameters obtained from data, and (3) Noisy-OR parameters assessed by the expert.

Our enhancement process replaced those elements of the CPT that had not enough data records to learn a distribution reliably, i.e., when the number of records found in the data set was lower than a *replacement threshold* (we specified this threshold as a percentage of all records in the data set). Figure 6 shows the relationship between the *replacement threshold* and the percentage of all CPT entries that were replaced by the Noisy-OR distributions. The percentage of replaced CPT entries seems to be directly proportional to the threshold.

Figure 7 shows the results for the three tested models for the window size of 1. It pictures the diagnostic accuracy of the models as a function of the *replacement threshold*. In addition we included the results for the single-disorder model. It appears that the highest accuracy was reached by the model whose CPTs were enhanced with the Noisy-OR parameters learned from data. The model with the expert-elicited Noisy-OR parameters performed consistently worse than the other models. The highest accuracy achieved by the models was 42%, 47%, and 43% for the CPT model, the data Noisy-OR model and the expert Noisy-OR model respectively.

Figure 8 shows the performance within each class for the three models. Again we observed that for almost each of the disorders the data Noisy-OR model performed better than the other models.

Subsequently we tested models in which all 27 Noisy-OR gates were quantified by Noisy-OR parameters only. The accuracy of the

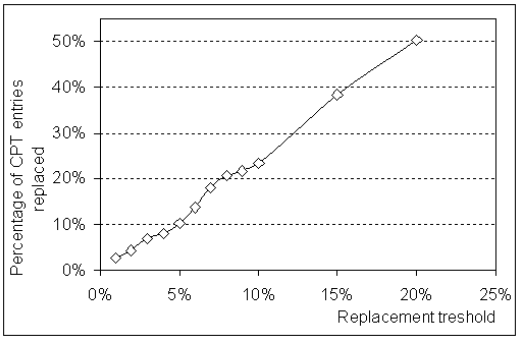

Figure 6. Percentage of conditional probability distribution entries replaced by Noisy-OR distributions as a function of the replacement threshold.

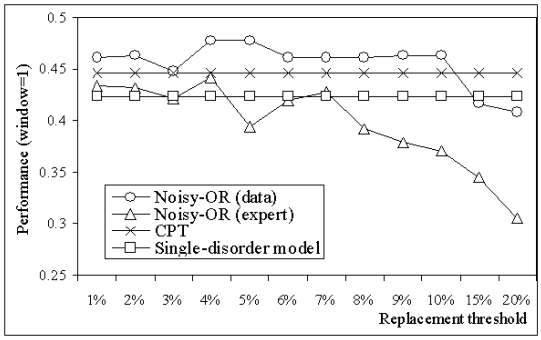

Figure 7. Diagnostic accuracy as a function of the replacement threshold, window=1.

Noisy-OR models was lower than the plain CPT model. The results for the window=1 were: 38% and 19% for the model with learned and expert-elicited Noisy-OR respectively.

7 DISCUSSION

The transformation of the model performed in order to prepare it for Noisy-OR gates has shown that Bayesian network models readily accommodate multiple-disorder diagnoses. It was relatively easy to derive the multiple-disorder version of the model from the existing single-disorder version. We estimate that the total time spent with the expert was less than 10 hours, one fourth of the original effort to build the network.

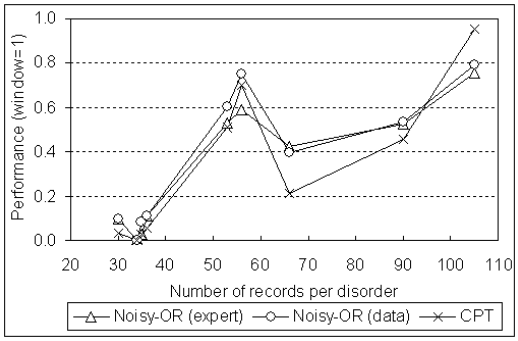

Figure 8. Diagnostic accuracy as a function of the number of disorder cases in the database (class size) for the CPT and two versions of the model with Noisy-OR parameters.

Diagnostic accuracy of the multiple-disorder model enhanced with the Noisy-OR parameters was around 6% better than the accuracy of the plain multiple-disorder model and 10% better than the single-disorder diagnosis model. This increase in accuracy has been obtained with very modest means — in addition to structuring the model so that it is suitable for Noisy-OR nodes, the only knowledge elicited from the expert and entered in the learning process was which interactions can be viewed as approximately Noisy-OR. This knowledge was straightforward to elicit. We have found that whenever combining expert knowledge with data, and whenever working with experts in general, it pays off generously to build models that are causal and reflect reality as much as possible, even if there are no immediate gains in accuracy.

We have also observed that the diagnostic accuracy of the model based on numbers elicited from the expert (as opposed to learned from data) was quite good for diseases with well understood risk factors and symptoms. The accuracy tended to be lower in case of those diseases whose mechanisms are not exactly known, for example hyperbilirubinemia, reactive hepatitis, or PBC, even if the number of records in the data set was very small.

Our plans for the future include expert verification of the probability distributions of those nodes that have several disorder nodes as parents. As we mentioned above, these parameters cannot be learned from our data and the arbitrary assumptions that we made in the learning process may have had a negative effect on the diagnostic performance of the system. We also plan to focus on disorder-to-disorder dependencies. This information is lacking from the database, so here again we will have to rely on expert judgment. As far as the use of Noisy-OR distributions to improve the quality of the model, we plan to investigate other ways of learning the parameters of Noisy-OR gates. Another question that we find worth pursuing is whether there are any properties of individual nodes that influence whether diagnostic accuracy of the model will be served by Noisy-OR or CPT used in individual cases.

ACKNOWLEDGMENTS

This research was supported by the Air Force Office of Scientific Research, grants F49620–97–1–0225 and F49620–00–1–0112, by the National Science Foundation under Faculty Early Career Development (CAREER) Program, grant IRI–9624629, by the Polish Committee for Scientific Research, grant 8T11E02917, by the Medical Centre of Postgraduate Education, Poland, grant 501–2–1–02–18/00, and by the Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, grant 16/ST/00. Our collaboration was enhanced by travel support from Poland to USA by NATO Collaborative Linkage Grant PST.CLG.976167.

We would like to thank anonymous reviewers and Dr. Javier Díez for their helpful comments on the paper. The HEPAR II model was created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory and available at `http://www2.sis.pitt.edu/~genie`.

REFERENCES

[1] Leon Bobrowski, 'HEPAR: Computer system for diagnosis support and data analysis', Prace IBIB 31, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland, (1992).

[2] Gregory F. Cooper and Edward Herskovits, 'A Bayesian method for the induction of probabilistic networks from data', *Machine Learning*, **9**(4), 309–347, (1992).
[3] F. Javier Díez, 'Parameter adjustment in Bayes networks. The generalized Noisy-OR gate', in *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pp. 99–105, Washington, D.C., (1993).
[4] F. Javier Díez and Marek J. Druzdzel. Canonical probabilistic models for knowledge engineering, 2000.
[5] F. Javier Díez, J. Mira, E. Iturralde, and Zubillaga S., 'DIAVAL, a Bayesian expert system for echocardiography', *Artificial Intelligence in Medicine*, **10**, 59–73, (1997).
[6] Nir Friedman and Moises Goldszmidt, 'Learning Bayesian networks with local structure', in *Learning and Inference in Graphical Models*, ed., Michael I. Jordan, 421–459, The MIT Press, Cambridge, MA, (1999).
[7] Max Henrion, 'Some practical issues in constructing belief networks', in *Uncertainty in Artificial Intelligence 3*, eds., L.N. Kanal, T.S. Levitt, and J.F. Lemmer, 161–173, Elsevier Science Publishers B.V., North Holland, (1989).
[8] B. Middleton, M.A. Shwe, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper, 'Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: II. evaluation of diagnostic performance', *Methods of Information in Medicine*, **30**(4), 256–267, (1991).
[9] A.W. Moore and M.S. Lee, 'Efficient algorithms for minimizing cross validation error', in *Proceedings of the 11th International Conference on Machine Learning*, San Francisco, (1994). Morgan Kaufmann.
[10] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk, 'A probabilistic causal model for diagnosis of liver disorders', in *Proceedings of the Seventh International Symposium on Intelligent Information Systems (IIS–98)*, pp. 379–387, Malbork, Poland, (June 15–19 1998).
[11] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk, 'A Bayesian network model for diagnosis of liver disorders', in *Proceedings of the Eleventh Conference on Biocybernetics and Biomedical Engineering*, volume 2, pp. 842–846, Warszawa, Poland, (December 2–4 1999).
[12] Agnieszka Oniśko, Marek J. Druzdzel, and Hanna Wasyluk, 'Extension of the Hepar II model to multiple-disorder diagnosis', in *Advances in Soft Computing*, Heidelberg, New York, (2000).
[13] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.
[14] Judea Pearl and Thomas S. Verma, 'A theory of inferred causation', in *KR–91, Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, eds., J.A. Allen, R. Fikes, and E. Sandewall, pp. 441–452, Cambridge, MA, (1991). Morgan Kaufmann Publishers, Inc., San Mateo, CA.
[15] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, and G.F. Cooper, 'Probabilistic diagnosis using a reformulation of the INTERNIST–1/QMR knowledge base: I. The probabilistic model and inference algorithms', *Methods of Information in Medicine*, **30**(4), 241–255, (1991).
[16] Peter Spirtes, Clark Glymour, and Richard Scheines, *Causation, Prediction, and Search*, Springer Verlag, New York, 1993.
[17] Sampath Srinivas, 'A generalization of the Noisy-OR model', in *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI–93)*, pp. 208–215, Washington, D.C., (1993).
[18] Hanna Wasyluk, 'The four year's experience with HEPAR-computer assisted diagnostic program', in *Proceedings of the Eighth World Congress on Medical Informatics (MEDINFO–95)*, pp. 1033–1034, Vancouver, BC, (July 23–27 1995).